EXHIBIT 4

[redacted]

UNITED STATES DISTRICT COURT
DISTRICT OF MASSACHUSETTS

SINGULAR COMPUTING LLC,)
)
Plaintiff,)
)
vs.) Case Nos.
) 1:19-cv-12551-FDS
GOOGLE LLC,)
)
Defendant.)
)

*** [redacted] ***

REMOTE VIDEO DEPOSITION OF
DR. SUNIL KHATRI, VOLUME II

DATE TAKEN: MARCH 24, 2023
REPORTED BY: RENEE HARRIS, CSR 14168, CCR, RPR
JOB NO. 5805112
PAGES: 350 - 699

Page 350

[redacted]

A. Okay. I'm at the claims section.

Q. Yeah, and you understand that it's Claim 7 that's being -- of the '156 patent that's being asserted in this case; right?

A. Yes, I'd like to just double-check that 05:32:11 to be 100 percent sure.

Q. Sure.

A. Yes, it is Claim 7 of the '156 patent; and that's the dependent claim which depends on 3, which depends on 2, which depend on 1. 05:32:33

Q. Claim 1 has been ruled invalid; correct?

MR. SEEVE: Objection. Calls for speculation.

THE WITNESS: I have no idea about that. I have no information to answer that question 05:32:50 either way.

BY MR. KAMBER:

Q. You participated in the IPR proceedings in this case; correct?

A. I -- I did -- you know, I did present -- 05:33:00 I mean, I did present -- I was deposed as well in the IPR proceedings. But the outcome of the IPR proceedings, I'm unaware of.

Q. You have no idea what the outcome of the IPR proceedings is? 05:33:17

[redacted]

That -- and that last part, the, "And produces unexpected results," is an important element that shouldn't be missed because -- because as Dr. Leeser says in her report that, oh, there's nothing surprising about -- you know, 06:25:04 there's no surprising results you might get by using additional execution units and so on. But this -- this line underscores that, no, there is a surprising result that you obtain because it does give unexpected results and these unexpected 06:25:21 results are what are described by, in the specification of the patent, from -- you know, from the columns I just recited to you which begin around, you know, column 16, lines 59, all the way through column 23 lines, I guess it was 34 or 06:25:39 something like this. There's explicit disclosure in the patent that -- of this unexpected result and of course this unexpected result is also described in Dr. Bates slides to Google as well.

So this line is actually underscoring 06:26:00 that unexpected result and it's important to not -- to not leave out that fragment of the line when you're citing it.

Q. Dr. Khatri, is it your belief that Dr. Bates was the first one to determine that 06:26:17

Page 578

[redacted]

using narrower bitwidths led to more parallelism?

MR. SEEVE: Objection -- objection. Mischaracterizes the report and the witness's prior testimony. Calls for a legal conclusion. 06:26:31

THE WITNESS: I'm going to say that, you know, Dr. Bates -- you know, the question you are asking is an incomplete question. Because what I'm -- my answer is, Dr. Bates was the first to describe multiple things and 06:26:45 among these things is, one, is that if you use low precision, high-dynamic execution units, A, you can fit more units, you know, in the same circuit area; B, that results in reduction -- I mean, B, that results in 06:27:00 dramatic improvement in performance at high precision, that's important, right. And that's the surprising result.

That's the part that -- you know, the totality of all these comments is what is 06:27:13 important, and that's the -- that's what is described in the patent in detail, as well as in Dr. Bates presentations to Google, which I cite on page 53 of my report where he shows -- you know, one thing he shows is 06:27:31

Page 579

[redacted]

that, you know, the approximate
floating-point units are much smaller; and
therefore, the next citation is the figure at
the bottom of page 53, which comes from his
slides, which says that, you know, that we 06:27:44
can have, you know, 100x more floating-point
units compared to the IEEE floating-point
units. That's the other thing he says. That
means you can squeeze in more floating-point
units in the same chip area. 06:27:59

But then the next slide which -- which I
cite in page 54 -- it shows that the software
can get, you know, 10,000x better speed and
power than the GPU, that's what the other
citation is. 06:28:15

And then finally, the other comment is
that, you know, he shows the -- you know,
which I -- which I show from his slide and
surprise No. 2, page 57 of my report, that
even though we have these low precision 06:28:28
units, you know, operating, you know, in
parallel, and I'll quote here this -- because
this is important.

It says, "The high precision CPU managing
low precision workers" -- that means LPHDR 06:28:42

[redacted]

execution units -- "can yield high precision results, like the CPU," completely unexpected result which is surprising and many Google engineers in their e-mail responses among each other expressed significant surprise at 06:28:57 this, and also it says, surprise No. 2 continues.

It says, "but with size, power, cost of the low precision hardware for varied tasks."

So not only are we going to get this, you 06:29:10 know, size, power, and cost comparable to -- of the low precision hardware but also this applies to many tasks. This is significant because this allows this idea to be used for many tasks and get tremendous speed up, you 06:29:26 know -- you know, with these low precision units, and the precision still is comparable to the "high precision CPU," completely surprising result.

BY MR. KAMBER: 06:29:39

Q. Those ideas were known before, though; right, Dr. Khatri?

MR. SEEVE: Objection -- objection. Mischaracterizes the witness's testimony. Argumentative. 06:29:46

Page 581

[redacted]

THE WITNESS: They were definitely not known and there is basically -- that's the inventiveness of the patent and that's also expressed in the e-mails the engineers exchanged among themselves once they saw the 06:30:02 second doc of Dr. Bates, and there was some significant praise that they expressed, significant surprise that they expressed.

There's multiple reasons why this was surprising to the community, because the 06:30:15 conventional wisdom -- in fact, the patent specification says this: The conventional wisdom is that if you want a high precision algorithmic output, you must use high precision execution units. 06:30:31

But this patent shows a completely surprising result, that if you want -- if you -- if you use low precision execution units and you can use many of them, for many applications, you can still get a high 06:30:44 precision output, which is significantly surprising and it's completely against the conventional wisdom in the field of -- in the field.

And there is disclosure in the patent, I 06:30:57

Page 582

[redacted]

can point you to it, where -- where Dr. Bates, the inventor sort of describes this.

BY MR. KAMBER:

Q. Let move to Exhibit 12 for a moment -- 06:31:09
Dr. Khatri, go to Exhibit 12, again, please.

MR. SEEVE: I'd like to point out that I think you just interrupted the witness's answer but --

MR. KAMBER: I disagree. 06:31:17

MR. SEEVE: Like you've done so many times.

MR. KAMBER: He was done with his answer. He was offering to --

THE WITNESS: Let me open Exhibit 12 real 06:31:28
quick. Excuse me. I have Exhibit 12 open in front of me.

BY MR. KAMBER:

Q. Go to page 5, please.

A. Can you give me, if you don't mind, the 06:31:44
title of that?

Q. "Format design trade-offs."

A. I see that slide.

Q. This slide shows, as Dr. Leeser was explaining, at this HPEC conference, that using 06:32:01

Page 583

[redacted]

which means you must use wider bitwidth,
which means you should get a high precision
functional unit, not a low precision
execution unit. The surprising -- let me
please finish -- the surprising and 06:33:19
significant aspect of the asserted patents is
that despite using narrower bitwidths, you
can get a high precision output which
doctor -- you know, which the patent
describes -- which Dr. Bates describes in 06:33:36
those paragraphs that I cited to you which I
think were paragraphs -- sorry, columns 14
through columns 23.

Those are concrete examples and concrete
experiments that Dr. Bates had conducted to 06:33:49
show that with narrower bitwidths, one can
still get high precision for many
applications, and that's completely
contradictory to this slide because this
slide says, to get high precision, you should 06:34:03
use wider bitwidths, because as the arrow
pointing to the right, saying wider bitwidths
gives rise to higher precision.

So Dr. Bates' observation, Dr. Bates'
patent and the asserted claims and the 06:34:17

[redacted]

asserted patents show this completely surprising phenomena, which the conventional wisdom, you know, simply didn't subscribe to, which is why, as I said, you know, [redacted]

[redacted]

BY MR. KAMBER:

Q. In that response, are you referring to low precision as construed by the Court or in some other sense?

MR. SEEVE: Objection. Mischaracterizes 06:35:11 the witness's testimony. Vague and ambiguous.

THE WITNESS: I don't understand your question, so -- when you say, when I'm referring to "precision," what do you mean? 06:35:22

[redacted]

that -- sorry.

For all the IPRs, because we are talking about IPRs, for all the IPRs that were -- that were filed, I'm not aware of the -- the legal paperwork that's filed back-and-forth 07:21:35 between Google and -- you know, and the PTAB or -- or Singular and the PTAB or Google and Singular.

I'm only aware of those documents that were made available to me for the analysis 07:21:51 that I needed to conduct which is purely technical and of course not legal because I'm not a lawyer.

So whatever documents were provided to me for my technical analyses, which were all 07:22:01 that I requested, those I reviewed. But subsequent to that, I'm not -- I'm not aware of the rulings or decisions that the PTAB has made about specific claim elements.

And to the extent that their analysis 07:22:18 defers from mine, I respectfully accept it but I disagree with it because I stand by my analysis.

BY MR. KAMBER:

Q. Do you have any understanding that 07:22:29